Detecting Similarities in Process Data generated with Computer-based Assessment Systems

Aleksandar Pejić

Computer-based Assessment Systems are nowadays widely adopted in the field of education. They are used to evaluate, measure and document the performance of students, from as soon as the early childhood (preschool), throughout primary, secondary and higher education. A prominent example is OECD's (Organisation for Economic Co-operation and Development) Programme for International Student Assessment (PISA), an international large-scale testing program on student performance. This study is conducted worldwide every three years, and measures the performance of 15-16 years old pupils. It focuses on mathematics, reading, science and problem-solving areas of assessment. While OECD regularly publishes the results of the study across the participating countries, it also makes publicly available the databases for each year the pupils took the test.

In PISA 2012 a computer-based assessment focusing on the problem-solving skills was introduced, that is particularly interesting for researchers in applied informatics and social sciences. The CBA system recorded not only the final result of the task at hand, but all the actions that the pupils performed in the CBA environment. Thus the resulting dataset is rich with information; it makes it possible to analyze the applied problem solving strategies [1], but also sets a number of challenges for the researchers. Adequate information extracting and aggregating methods have to be identified. To obtain proper information for complex analyzes, as well as to determine its meaning in the domain of social sciences, experts from the field have to be consulted.

In this work we are investigating the process of extracting, aggregating and finding similarities between sequences of actions recorded for a PISA 2012 problem-solving item. Identifying and assembling the feature sequence for each test taker is a prerequisite for further researching and explaining cognitive performance.

One possible approach is to utilize natural language processing (NLP) methodologies. A technique for analyzing problem-solving process data based on N-grams was developed on the premise of similar structure among action sequences and word sequences in natural language [2], following the conclusion that the evolution and use of sequential models is closely related to the statistical modeling of texts [3]. Another important aspect taken from text categorization is identifying the key features, as well as disregarding the insignificant features, by applying a corresponding strategy, known as feature selection, which provides basic means for classification [4].

The described technique gives us a feature sequence for each test taker. We are following up by investigating a method for clustering the test takers based on the similarity of their sequences. In natural language processing and information retrieval a text document is often represented as a simplified multiset of its words. This model is known as bag-of-words [5], or bag-of-features in our case. After assembling a multiset of selected features, we can represent each test takers' feature sequence as a vector. These vectors are then quantified and grouped using k-means clustering. For data processing and clustering the Weka software was used (Waikato Environment for Knowledge Analysis).

References

[1] S. Greiff, S. WĂźstenberg, F. Avvisati. Computer-generated log-file analyses as a window into studentsḿinds? A showcase study based on the PISA 2012 assessment of problem solving. Elsevier Computers & Education, 91, 92-105, 2015

[2] Q. He, M. von Davier. Identifying Feature Sequences from Process Data in Problem-Solving Items with N-Grams. Springer Proceedings in Mathematics & Statistics, 140, 179-196, 2014

[3] G. A. Fink. Markov models for pattern recognition. Berlin, Germany, Springer, 2008

[4] S. Li, R. Xia, C. Zong & C. Huang. A framework of feature selection methods for text categorization. In Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP, 692-700, 2009

[5] G. Salton, M.J. McGill. Introduction to modern information retrieval. McGraw-Hill, 1983